**An examination of how translation experience influences cognitive load in a real-world situation using EEG**

Matthias Kobi1, Michael Boos1, Stefan Elmer1\* & Lutz Jäncke1,2\*

1Division Neuropsychology, Department of Psychology, University of Zurich, Zurich, Switzerland.

2University Research Priority Program (URPP) “Dynamics of Healthy Aging”, University of Zurich, Zurich, Switzerland.

\* Shared last authorship

Keywords:

Cognitive load, translation, language expertise, EEG

E-mail addresses:

MK: [matthias.kobi@psychologie.uzh.ch](mailto:matthias.kobi@psychologie.uzh.ch)

MB: m.boos@psychologie.uzh.ch

SE: s.elmer@psychologie.uzh.ch  
LJ: l.jaencke@psychologie.uzh.ch

Correspondence to:  
Matthias Kobi or Stefan Elmer  
Institute of Psychology  
Division Neuropsychology  
University of Zurich  
Binzmühlestrasse 14/25  
8050 Zurich, Switzerland

Date: 27 March, 2022

**Abstract**

**Introduction**

*Importance of bilingualism*

*Neurophysiological evidence of translation*

*How does expertise and experience affect the translating brain*

*Evidence from non-standard language input (ELF vs EdE)*

*Cognitive load*

most of the studies focusing on mental workload were using traffic-related scenarios or working memory tasks, less work has been done regarding language processing and workload

*- Cognitive load theory*

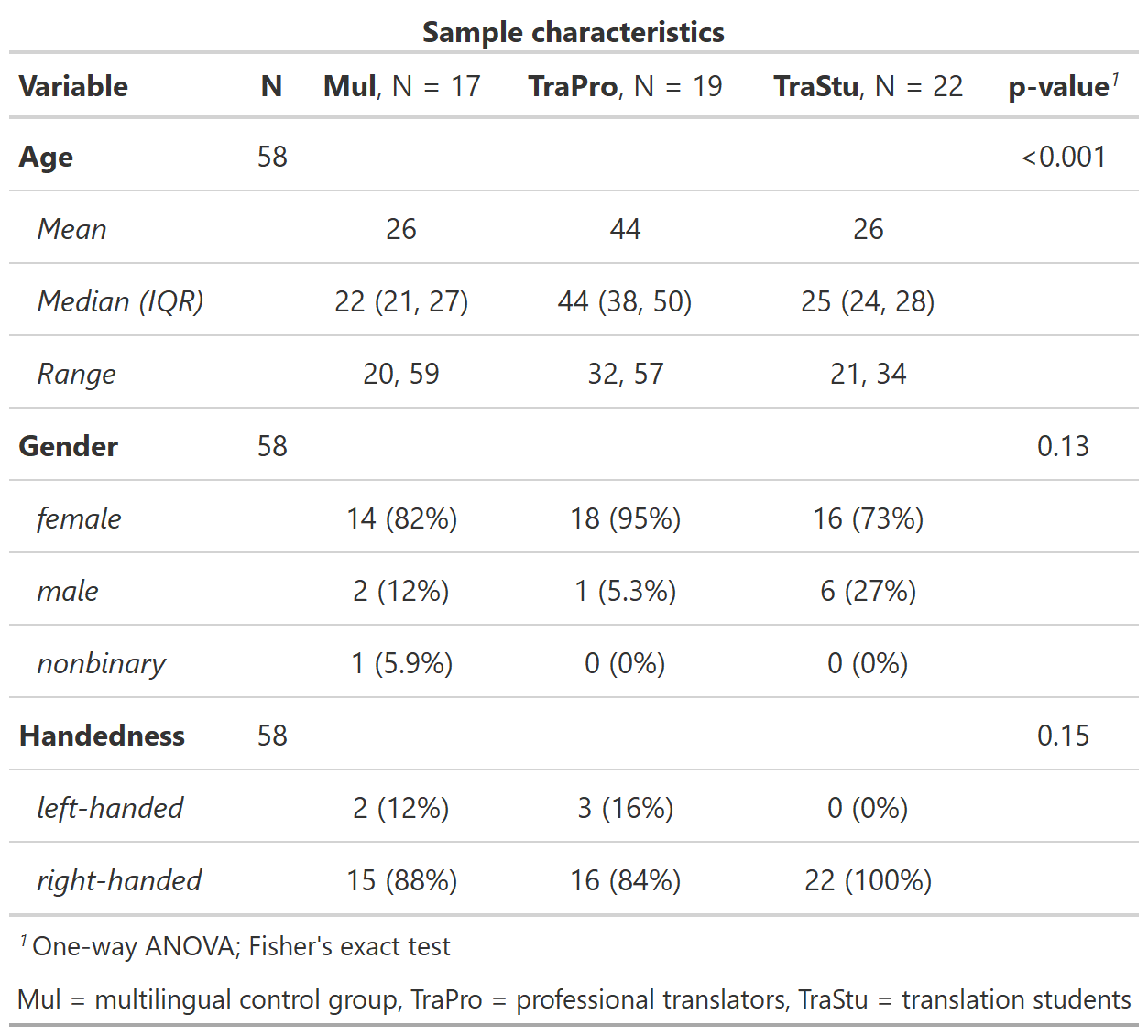
*- Measuring cognitive load with EEG*

*Hypothesis regarding our research*

**Methods**

**Participants**

We collected data from 72 native German (L1) participants (62 females, one nonbinary) with a professional English (L2) background. We recruited three different groups: professional *written* translators (TraPro, N = 19), trainee translators (TraStu, N = 22), and a multilingual control group (Mul, N = 17). Members of the multilingual control group were English language and literature studies students or teachers of English as a foreign language from high schools. All participants were required to use L2 in their daily routine, and the primary direction of translating for professional and student translators was from L2 to L1. Since we recruited participants with varying levels of professional experience, groups could not be matched regarding their age (Table X). However, they did not differ in respect of gender or handedness (Annett, 1970). All participants had a normal or corrected-to-normal vision. Two participants reported using medicaments for diabetes, two for high blood pressure, and two reported concussions that occurred longer than five years before testing. The experiment lasted approximately four hours and was rewarded with cash. Fourteen participants were excluded from the analysis because of failure to follow the instructions of the experiment (N = 6), medication (N = 4, anti-depressants or Ritalin), and noisy or missing data (N = 4). Thus, we analyzed data from 58 participants. The study was carried out according to the principles in the declaration of Helsinki and approved by the Swiss National Science Foundation ethics committee.



*Table 1: Sample characteristics*

**Psychometrics and questionnaires**

Every participant completed a short English language test ([https://www.sprachtest.de/ einstufungstest-englisch](https://www.sprachtest.de/einstufungstest-englisch)) to assess L2 proficiency. This online procedure lasted about 15 minutes and consisted of 13 vocabulary, grammar, listening, and reading comprehension questions. The maximum score of the test was 40. Furthermore, we collected data on the age of L2 acquisition and experience in translating and interpreting (cumulative training hours and cumulative training hours per day since the age of 17) in a language background questionnaire. To assess working memory capacity, participants completed both a visual and an auditory 3-back task comprising 60 letter stimuli, of which 20 were target stimuli. The order of the tasks was pseudorandomized across the groups. We analyzed N-back data using d-primes (d’). D-primes were calculated as the difference between the z-transformed hit rate and false alarm rate (Hautus et al., 2021). Additionally, we evaluated the cognitive capabilities using a short version of the WAIS (Wechsler Adult Intelligence Scale) test battery (Waldmann, 2008). This short version was composed of four subtests: number-symbol associations, detection of commonalities, the mosaic test, and digit span forward and backward. Using the standardized T-values, the four subtests sensitively reflect general intellectual abilities (Waldmann, 2008).

**Stimulus material**

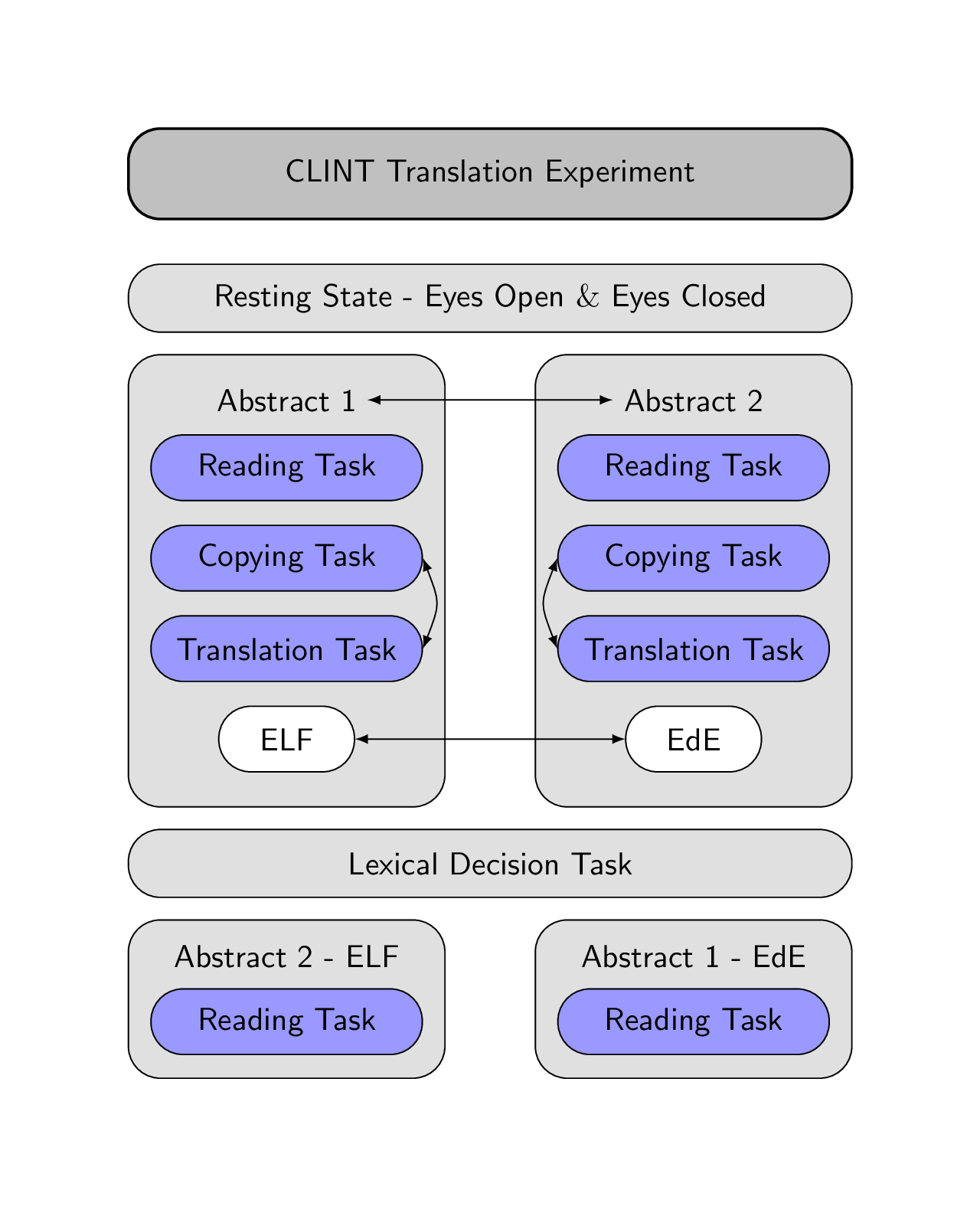
In this experiment, we used two original English abstracts submitted to conferences. Because the authors of the abstracts were non-native English writers, their original texts were regarded as ELF stimuli. In the next step, the two abstracts were processed into an edited-to-standard English (EdE) version by professional translators of the Zurich University of Applied Sciences (ZHAW). As few changes as possible were made to keep the text as close to the original while generating grammatically correct sentences and overall better readability. This translation procedure resulted in four different text stimuli: text 1 (ELF, original), text 1 (EdE), text 2 (ELF, original), and text 2 (EdE).

*- where are the abstracs from*

*- how many sentences per version*

*- How many words per version*

**Experimental procedure**

****

*Figure 1: Experimental design. Arrows indicate randomizations in the experiment.*

First, participants completed an EEG resting-state eyes open, and eyes closed condition for three minutes each. Second, participants started with the reading task, followed by the copying and translation task for the first abstract. In the reading task, the text was presented sentence by sentence, and participants could read at a self-paced speed through a button press. Subsequently, we asked the participants how difficult they thought this task was, and the answer was collected through mouse press on a 10 cm horizontal bar (easy on the left, difficult on the right). Additionally, we asked five control questions to check if participants had read the text. Those questions were multiple-choice, and participants had to choose one of three answer possibilities by pressing a key on the keyboard. Third, in the copying task, participants were asked to copy the presented sentences, and therefore, the generated output was English (L2). After completing a sentence, participants could move on to the following sentence by pressing “Enter”. Fourth, in the translation task, the presented sentences were then translated to German (L1), reflecting a forward translation (FT). Again, after completion of a sentence, participants had to press “Enter” to continue with the following sentence. After the translation task, we asked the participants again how difficult they thought this task was, while collecting the answers on a 10 cm horizontal bar.

In all tasks, the words of the presented sentence were separated with double spacing and double lines. In the reading, copying and translation task, the sentences that had to be processed were displayed in the upper part of the monitor, while the participants’ answers in the copying and translation task were presented in the lower part. The duration of the reading task differed based on the self-paced reading of participants. However, the copying and translating task was limited to five minutes each. After working on the first abstract, participants continued using the same procedure with the second abstract.

In the experiment, we randomized the order of the abstracts (text 1, text 2), the version (ELF, EdE), and the copying and translation task across participants, indicated by the arrows in Figure X. Therefore. each participant processed an abstract only in one version but not in the other. If the first abstract was in ELF, the second was in EdE and vice versa. Since the copying and translation task duration was limited, participants did not process the whole text but always started from the beginning and worked through the text sentence by sentence. However, it was made sure that no sentence was used twice in the copying and translation task.

After processing the two abstracts, participants completed a lexical decision task and then had to reread the abstracts in the other versions. However, we did not include data from those two conditions in the analyses. At the beginning of the experiment, instructions for the task were presented on the computer screen, and to become confident with the keyboard, participants had to copy a sentence that contained all possible special symbols from the abstract.

**Data acquisition**

After written informed consent, participants completed all psychometric measurements. Afterward, they were prepared for behavioral and EEG data acquisition, which took place in a light-dimmed Faraday cage where the participants were seated approximately 70 cm in front of a 24-inch monitor. The participants were instructed to relax and stay as still as possible during the EEG measurements. The experiment was programmed in MATLAB 2016b using the Psychophysics Toolbox Version 3 extension (Kleiner et al., 2007) for behavioral data acquisition. We recorded high-density EEG data at a sampling rate of 500 Hz with a bandpass filter of 0.1-100 Hz using the EGI 300 Geodesic EEG system with a 128-channel HydroCel Geodesic Sensor Net (HCGSN) (Electrical Geodesics, Eugene, Oregon). Before recording, each electrode was double-checked to ensure good contact on the scalp, and impedances were kept below 40 kOhm. This procedure was repeated after the EEG resting state, the processing of abstracts 1 and 2, and the lexical decision task. The recording reference electrode was Cz.

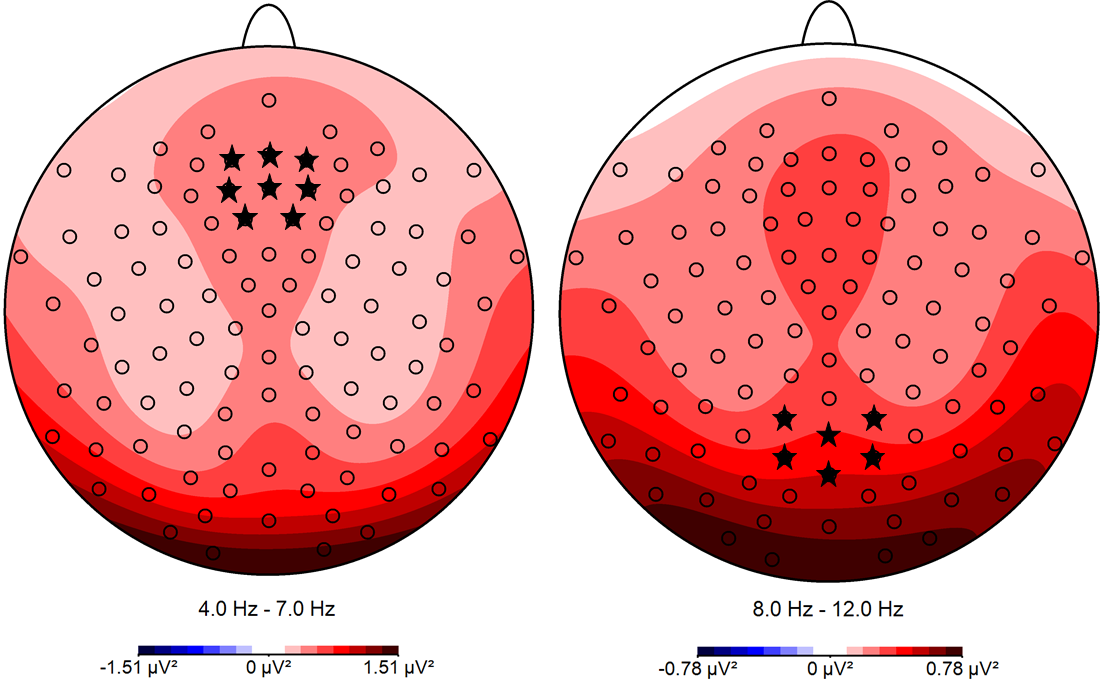
**Behavioral data processing**

The preprocessing of the behavioral data was done using R (version 3.6.3, <https://www.r-project.org/>). For the reading task, we evaluated the percentage of the correct answered control questions per text as well as the average reading duration per sentence, which was adjusted for the different lengths of the texts. Furthermore, we analyzed the perceived difficulty as the distance in cm of the mouse click from 0 (easy) for both the reading and translation tasks. Therefore, higher values indicate a more pronounced perceived difficulty of the task. Regarding the copying and translation tasks, we evaluated the total amount of chars typed during the 5 minutes, as well as the number of chars typed if deletions were subtracted (chars end version). Those variables can be regarded as a measure of the efficiency in the copying and translation task (Quelle). Furthermore, we retrieved the percentage of deletions, which refers to pressing the “backslash” on the keyboard relative to the total number of chars typed for both tasks. Finally, we analyzed the output generated by each participant in the translation task by rating the fluency (0: incomprehensible – 5: flawless German) and the accuracy ( 0: no meaning – 5: all meaning) per sentence (Koehn & Monz, 2006). For the rating, we fully randomized the sentences from both texts and conditions and all participants. Three independent raters (*language experts from the IUED Institute of Translation and Interpreting of the ZHAW*) rated the fluency first and, subsequently, the accuracy of all sentences. For the accuracy rating, the translation output was compared to a reference translation provided by the IUED. Then, the sentence ratings were averaged per condition (ELF vs. EdE) and both texts to calculate an intraclass correlation coefficient (ICC) using the irr package (Version 0.84.1, <https://cran.r-project.org/web/packages/irr/>) in R. Applying a 2-way mixed-effects model of the type “consistency” and a mean-rating (k=3) revealed a ICC(C,3) = 0.575 (95%-confidence interval = 0.421 – 0.694) for the fluency rating and a ICC(C,3) = 0.909 (95%-confidence interval = 0.875 – 0.934) for the accuracy rating. The ICC for the fluency rating likely reflects moderate reliability, whereas the ICC for the accuracy rating reflects excellent reliability (Koo & Li, 2016). Finally, we averaged the three raters to generate a mean rating score for fluency and accuracy further used in the statistical analyses.

**EEG data processing**

The data was processed using MATLAB (2018), EEGLAB (version 2021\_0, Delorme & Makeig, 2004), and Brain Vision Analyzer (version 2.2.0, BrainProducts, Munich, Germany). For EEG data preprocessing, we used the Automagic toolbox implemented in MATLAB (v.2.5, Pedroni, Bahreini, & Langer, 2019), which is a pipeline for automatic EEG data cleaning. First, the number of EEG channels was reduced to 105 by discarding channels lying on the neck and face. Second, we applied the PREP pipeline for bad channel detection with the minimum variance set to 1. Third, we used the ICLabel approach with a temporary 2 Hz high-pass filter for artifact correction to remove muscle, eye, heart, and channel noise components with a probability threshold higher than 0.8. Fourth, we selected eleven frontal electrodes for the electrooculogram (EOG) regression. Fifth, power line noise was removed using the ZapLine method (de Cheveigné, 2020), eliminating five components. Sixth, we applied a 0.1 Hz high-pass and a 30 Hz low-pass filter. Finally, bad channels were reconstructed through spherical interpolation, and we applied the detrending algorithm to remove slow drifts. This procedure was applied independently for each task to avoid preprocessing unnecessary noise during pauses.

Further preprocessing of the EEG data was executed in the Brain Vision Analyser. First, we re-referenced the data to an average reference montage, and segmented the EEG into the different task segments. Second, we used an automatic raw data inspection to mark bad time windows indicating remaining artifacts that were not removed by Automagic. Third, we divided the data into segments of 2 s length without overlaps, in which data segments marked as bad were skipped. Forth, a fast Fourier transform (FFT) with a Hanning window (Length = 10%) was applied to all remaining segments. The resulting transforms were averaged per participant and condition before exporting the power values for the theta (bandwidth from 4 to 7 Hz) and alpha (bandwidth from 8 to 12 Hz) band. Based on the voltage distribution of the grand average across the reading, copying, and translation tasks, we analyzed theta power at a frontal (E4, E5, E10, E11, E12, E16, E18, and E19) and alpha power at a parietal (E61, E62, E67, E72, E77, and E78) electrode pool (see Figure X for electrode positions). Finally, we averaged the power per pool and frequency band for statistical analysis.

**

*Figure X: Topographical voltage distribution maps for theta and alpha band across reading, copying and translation tasks, and all participants. The channels selected for analyses are marked with \*.*

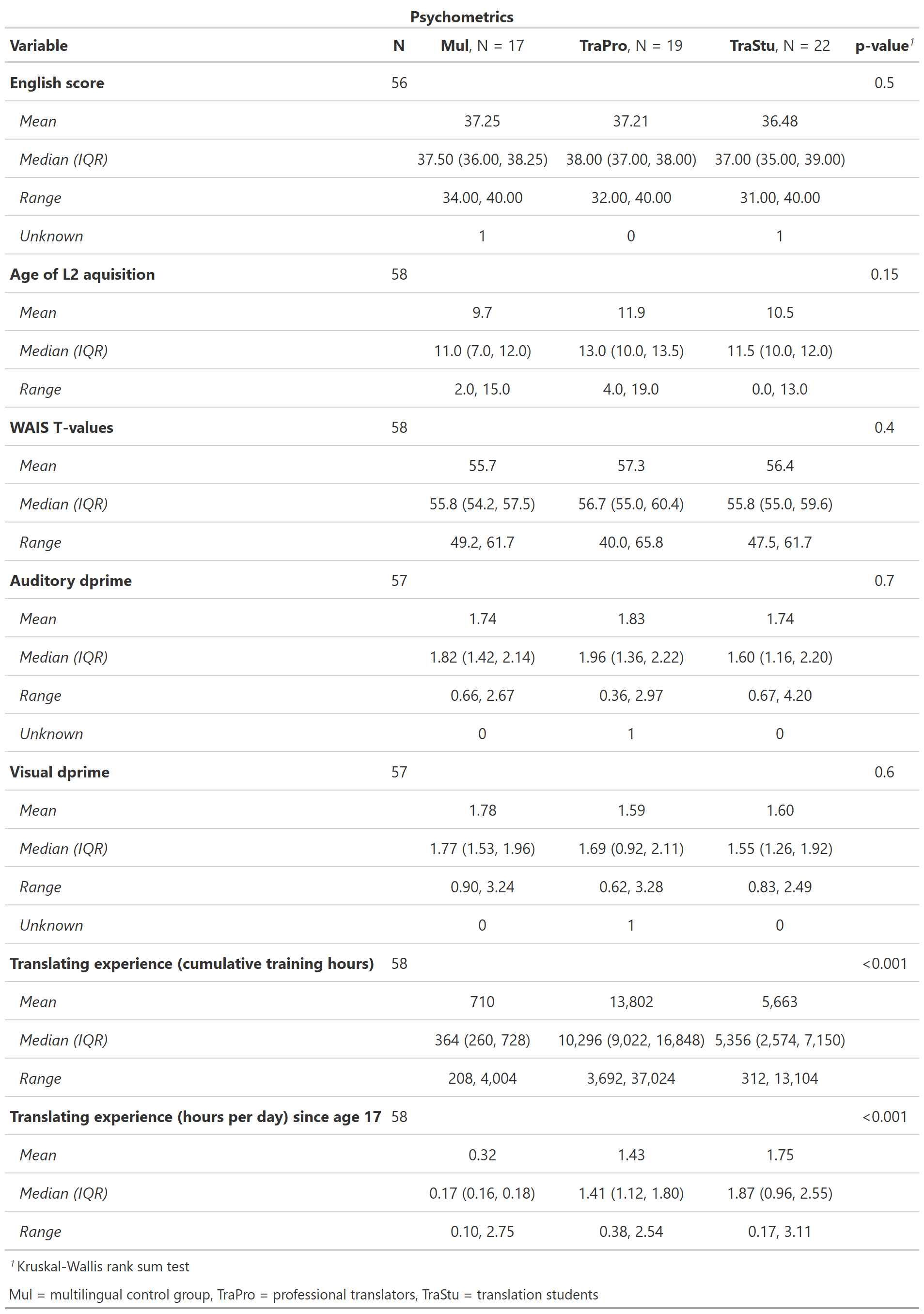
**Statistical analyses**

All statistical analyses were performed using Linear Mixed Models (LMM) implemented in the lme4 package (Version 1.1-23, <https://cran.r-project.org/web/packages/lme4/>) in R. For model-fitting, we used a bottom-up strategy starting with the null model and added fixed effects for our target variables. In general, we used three levels for task (reading, copying, and translation task), two levels for text (text1 and text2), two levels for condition (EdE and ELF) as well as three levels for group (TraPro, TraStu, and Mul).

**Results**

**Psychometrics and questionnaires.**

Our groups did not differ regarding English score, age of L2 acquisition, WAIS T-values, auditory d’, and visual d’ (Table X). However, as expected by our recruitment, our groups differed regarding cumulative training hours in translating and interpreting (TraPro: M = 13’802, TraStu: M = 5’663, Mul: M = 710, F(2,55) = 30.895, *p* < 0.001, η2G = 0.529), as well as in the cumulative training hours per day since the age of 17 (TraPro: M = 1.43, TraStu: M = 1.75, Mul: M = 0.32, F(2,55) = 18.858, *p* < 0.001, η2G = 0.407).

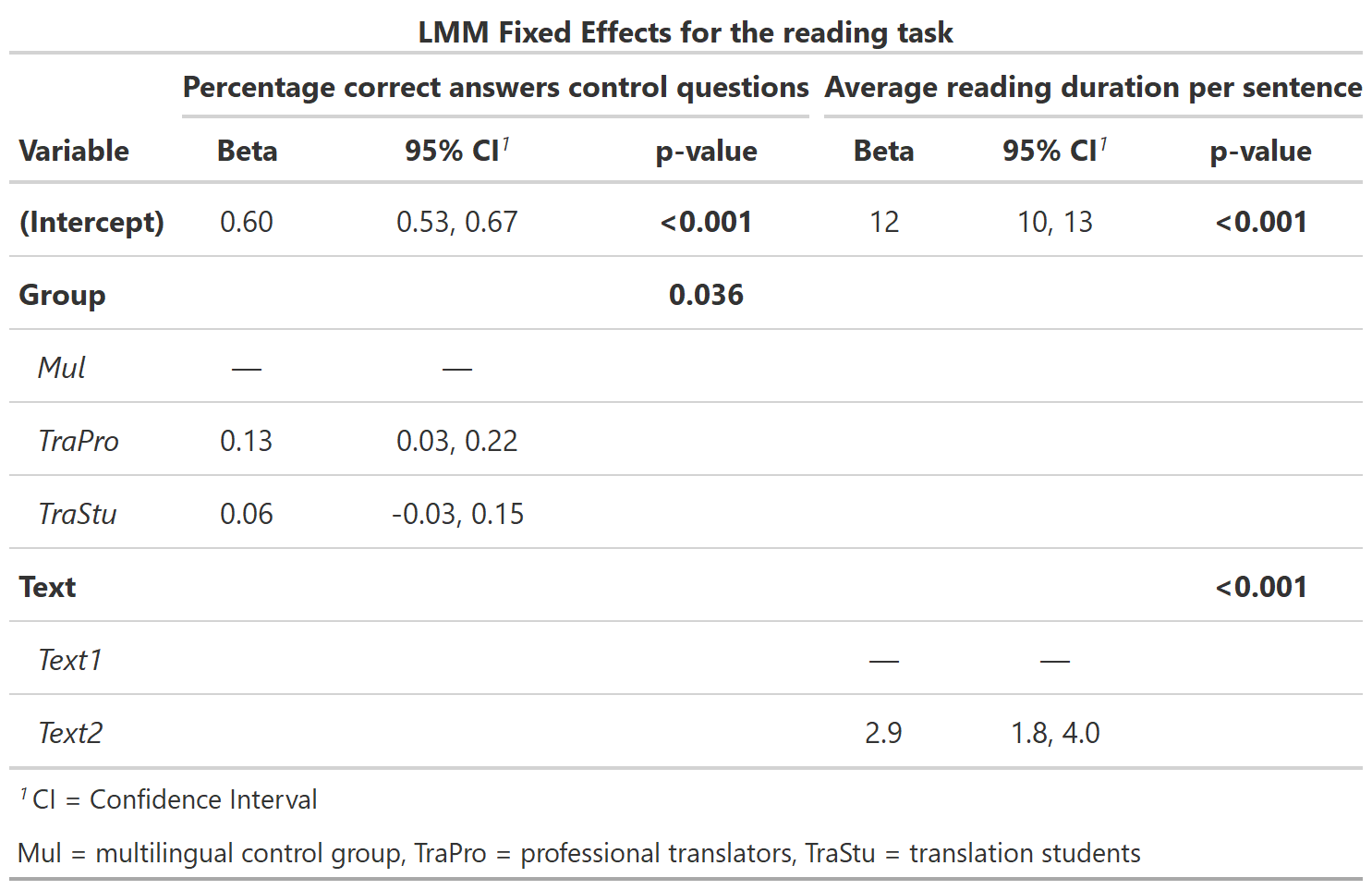
****

*Table X: Results of the psychometrics and questionnaires.*

**Behavioral results**

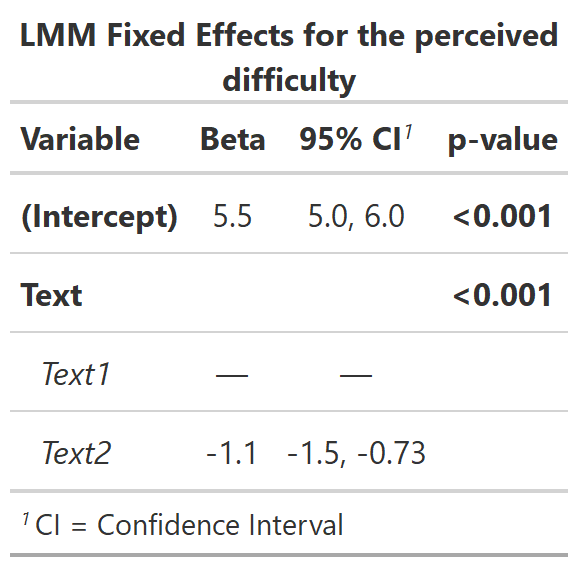
**Reading task**

For the analysis of the average reading duration per sentence, including a fixed effect for text significantly improved model fit (∆Χ2 = 23.559, *p* < 0.001), indicating a longer reading duration for the second text (*β* = 2.900) compared to the first text. Introducing a fixed effect for condition (∆Χ2 = 0.072, *p =* 0.789) as well as group (∆Χ2 = 3.540, *p =* 0.170) did not significantly improved model fit. Thus, the average reading duration per sentence was best predicted by text. For the analysis of the percentage of correct answers to the control questions, including a fixed effect for text (∆Χ2 = 1.89, *p =* 0.169) as well as condition (∆Χ2 = 0.074, *p =* 0.785) did not significantly improved model fit. Introducing a fixed effect for group significantly improved model fit (∆Χ2 = 6.619, *p* = 0.037), indicating a more accurate responses for the group TraPro (*β* = 0.126) and TraStu (*β* = 0.059) compared to the multilingual control group. Thus, the percentage of correct answers to the control questions was best predicted by group. All LMM fixed effects of behavioral measurements in the reading task are summarized in Table X.

****

**Perceived difficulty of the reading and translation task**

For the analysis of the perceived difficulty of the reading and translation task, including a fixed effect for task (∆Χ2 = 0.006, *p =* 0.941) did not significantly improve model fit. Introducing a fixed effect for text significantly improved model fit (∆Χ2 = 28.166, *p* < 0.001), indicating that the second abstract was perceived as less difficult (*β* = -1.128) compared to the first abstract. Adding a fixed effect for condition (∆Χ2 = 0.001, *p =* 0.993) as well as group (∆Χ2 = 3.125, *p =* 0.230) did not significantly improve model fit. Thus, the perceived difficulty of the reading and translation task was best predicted by text. The summary of all LMM fixed effects of the perceived difficulty is summarized in Table X.

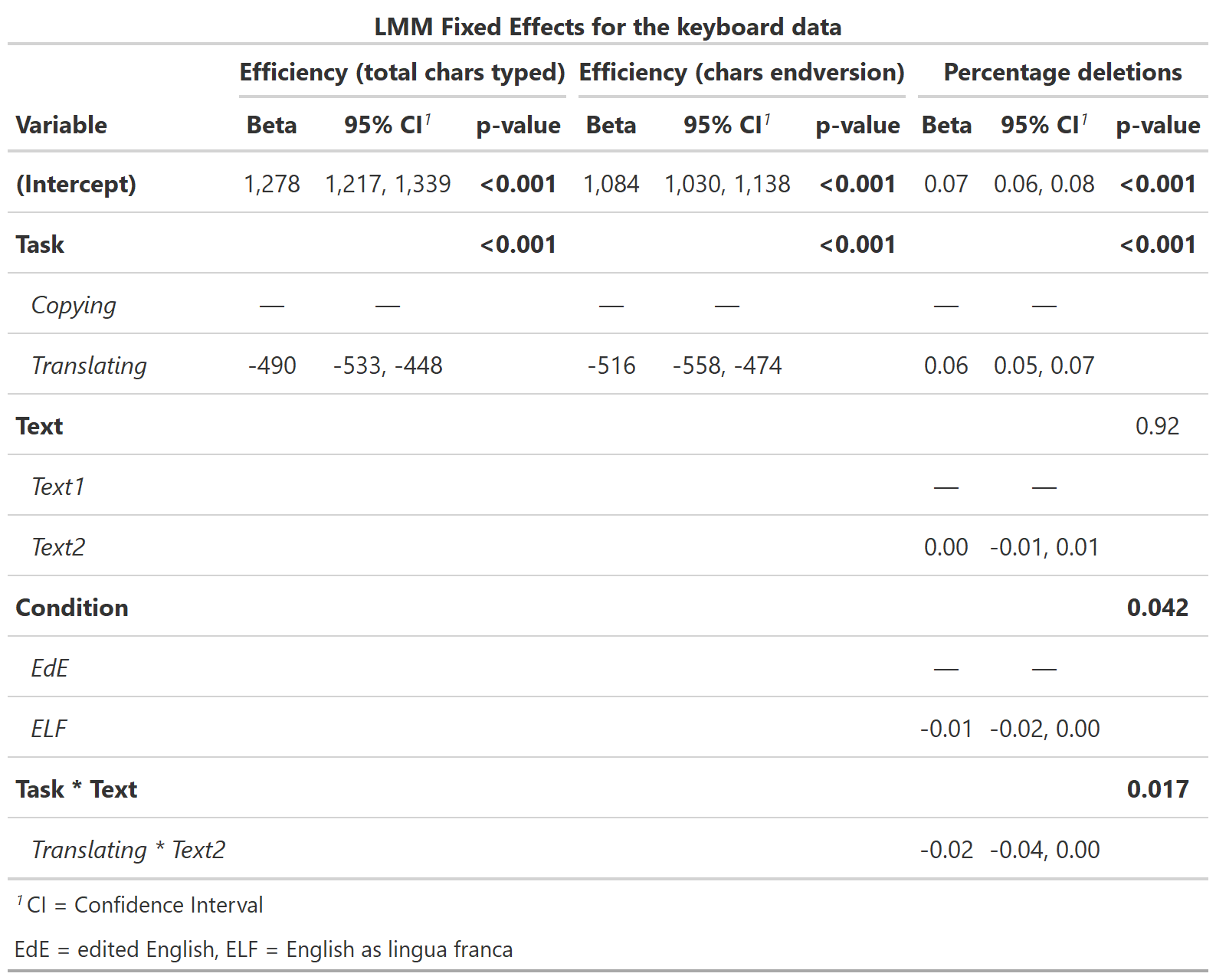
****

**Keyboard data**

For the analysis of the total number of chars typed in the copying and translation task, including a fixed effect for task significantly improved model fit (∆Χ2 = 239.63, *p <* 0.001), indicating that the total number of chars typed was lower in the translating task (*β* = -490.28) compared to the copying task. Introducing a fixed effect for text (∆Χ2 = 0.299, *p =* 0.585), condition (∆Χ2 = 1.062, *p =* 0.303) as well as group (∆Χ2 = 1.192, *p =* 0.551) did not significantly improved model fit. Thus, the total number of chars typed in the copying and translation task was best predicted by task.

For the analysis of the number of chars in the end version of the copying and translation task, including a fixed effect for task significantly improved model fit (∆Χ2 = 253.8, *p <* 0.001), indicating that the number of chars in the end version was lower in the translating task (*β* = -516.13) compared to the copying task. Adding a fixed effect for text (∆Χ2 = 0.271, *p =* 0.603), condition (∆Χ2 = 1.848, *p =* 0.174) as well as group (∆Χ2 = 1.469, *p =* 0.480) did not significantly improved model fit. Thus, the number of chars in the end version of the copying and translation task was best predicted by task.

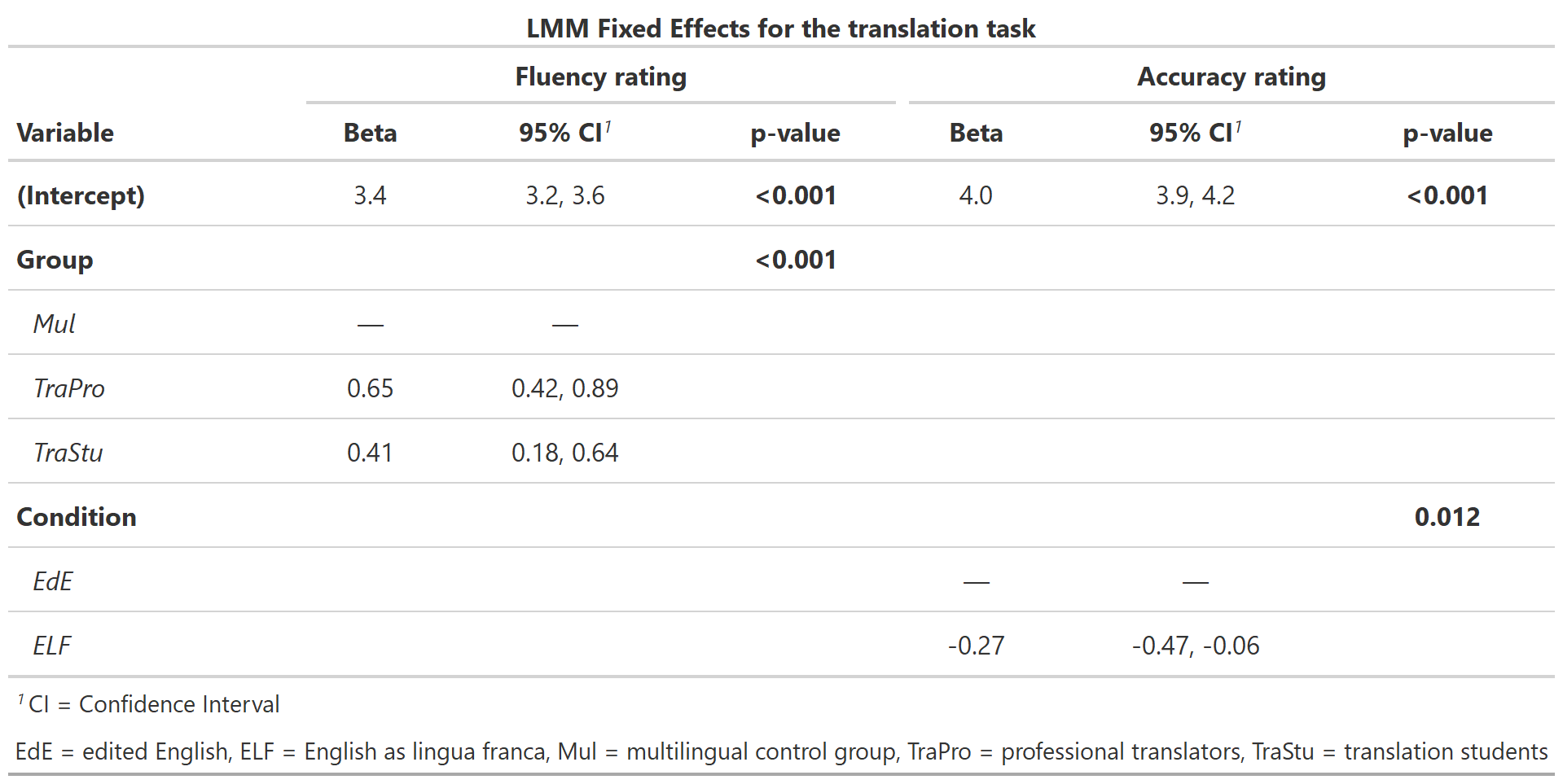
For the analysis of the percentage of deletions in the copying and translation task, including a fixed effect for task significantly improved model fit (∆Χ2 = 93.565, *p <* 0.001), indicating that the percentage of deletions was higher in the translating task (*β* = 0.052) compared to the copying task. Introducing a fixed effect for text significantly improved model fit (∆Χ2 = 6.110, *p =* 0.013), revealing that the percentage of deletions was lower for the second abstract (*β* = -0.011) compared to the first abstract. Adding a fixed effect for condition significantly improved model fit (∆Χ2 = 4.067, *p =* 0.044) showing that the percentage of deletions was lower for the ELF version (*β* = -0.001) compared to the EdE version. Including a fixed effect for group (∆Χ2 = 1.254, *p =* 0.534) did not significantly improved model fit. However, modeling an interaction between task and text significantly improved model fit (∆Χ2 = 5.745, *p =* 0.017), whereas the interaction between text and condition (∆Χ2 = 0.116, *p =* 0.733), and task and condition did not (∆Χ2 = 0.069, *p =* 0.793). The interaction between task and text reflects a lower difference in the percentage of deletions between the copying and translation task for the second abstract (*β* = -0.021). Thus, the percentage of deletions in the copying and translation task was best predicted by task, text, condition, and the interaction between task and text. All LMM fixed effects of the keyboard data are summarized in Table X.

****

**Translation task**

For the analysis of the fluency rating, including a fixed effect for text (∆Χ2 = 0.018, *p =* 0.893) as well as condition (∆Χ2 = 0.035, *p =* 0.852) did not significantly improved model fit. Introducing a fixed effect for group significantly improved model fit (∆Χ2 = 25.768, *p <* 0.001), indicating that the translations of the group TraPro (*β* = 0.65) and TraStu (*β* = 0.41) were rated to be more fluent compared to the Mul group. Thus, the fluency rating was best predicted by group.

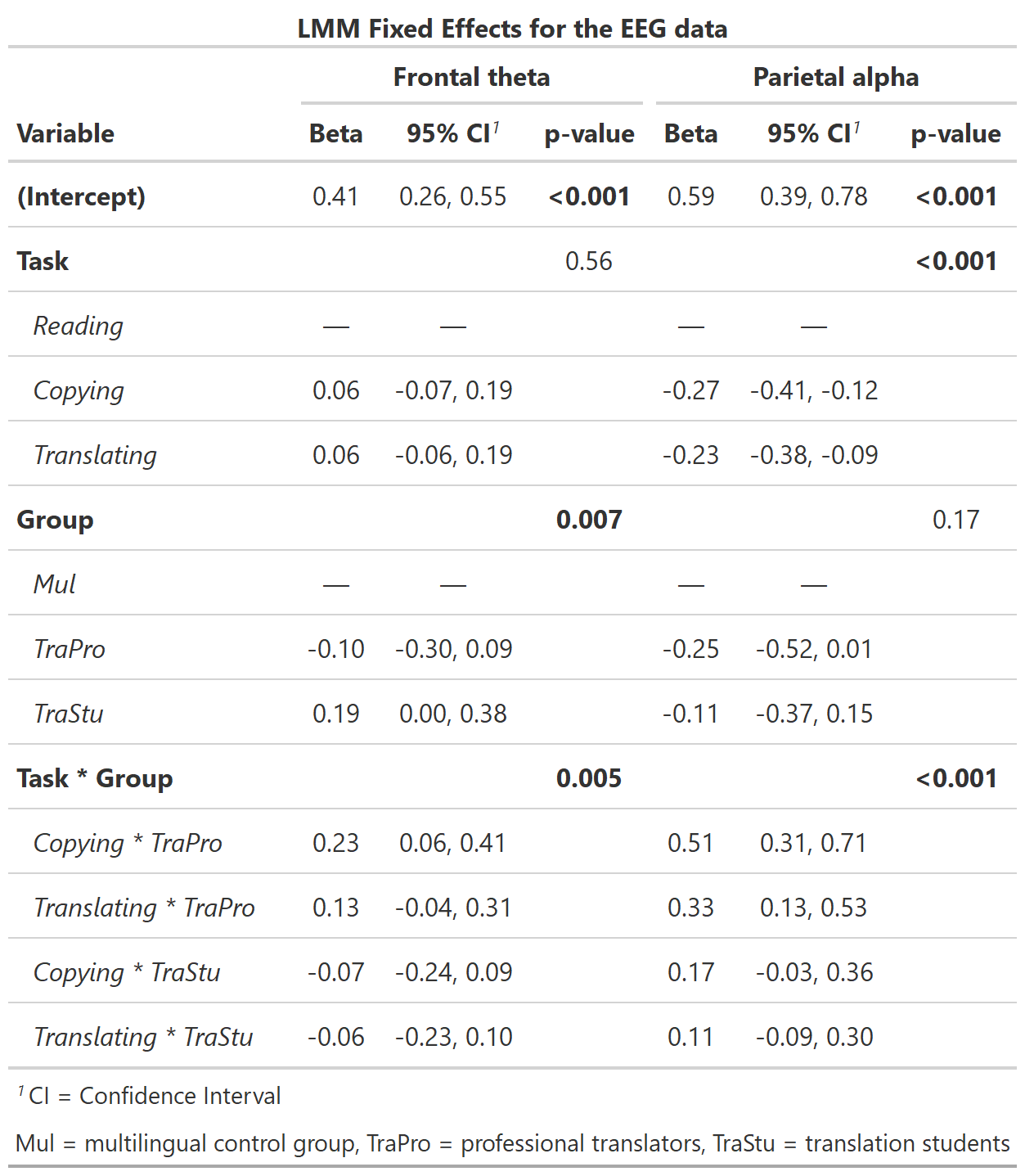
For the analysis of the accuracy rating, including a fixed effect for text (∆Χ2 = 0.171, *p =* 0.679) did not significantly improve model fit. Introducing a fixed effect for condition significantly improved model fit (∆Χ2 = 6.314, *p =* 0.012), indicating that the translations of the ELF version (*β* = -0.266) were rated to be less accurate compared to the EdE version. Adding a fixed effect for group (∆Χ2 = 1.850, *p =* 0.396) did not significantly improved model fit. Thus, the accuracy rating was best predicted by condition. All LMM fixed effects of the translation task are summarized in Table X.

****

**EEG results**

For the analysis of the frontal theta, including a fixed effect for task significantly improved model fit (∆Χ2 = 9.646, *p =* 0.008), indicating higher theta activity in the copying task (*β* = 0.106) and translating task (*β* = 0.081) compared to the reading task. Introducing a fixed effect for text (∆Χ2 = 1.408, *p =* 0.235), condition (∆Χ2 = 0.310, *p =* 0.577) as well as group (∆Χ2 = 3.640, *p =* 0.162) did not significantly improved model fit. However, adding an interaction between task and group significantly improved model fit (∆Χ2 = 14.623, *p =* 0.006). In the reading task, the group of TraPro revealed lower theta band activity (*β* = -0.105) compared to the Mul group, whereas the group of TraStu revealed higher theta band activity (*β* = 0.186). In the copying task, the group of TraPro (*β* = 0.126) and TraStu (*β* = 0.111) were reflected by higher theta band activity compared to the Mul group. In the translating task, the group of TraPro (*β* = 0.030) and TraStu (0.123) showed higher theta band activity compared to the Mul group. Thus, the frontal theta activity was best predicted by task, group, and the interaction between group and task.

For the analysis of the parietal alpha, including a fixed effect for fixed effect for task (∆Χ2 = 4.223, *p =* 0.121), text (∆Χ2 = 0.650, *p =* 0.420), condition (∆Χ2 = 0.544, *p =* 0.461) as well as group (∆Χ2 = 0.174, *p =* 0.917) did not significantly improved model fit. However, adding an interaction between task and group significantly improved model fit (∆Χ2 = 26.611, *p <* 0.001) compared to a model with only main effects for task and group. In the reading task, the group of TraPro (*β* = -0.252) and TraStu (*β* = -0.107) revealed lower alpha band activity compared to the Mul group. In the copying task, the group of TraPro (*β* = 0.260) and TraStu (*β* = 0.058) were reflected by higher alpha band activity compared to the Mul group. In the translating task, the group of TraPro (*β* = 0.080) showed higher alpha band activity compared to the Mul group, whereas the group of TraStu revealed lower alpha band activity (*β* = -0.001). Thus, the parietal alpha activity was best predicted by task, group, and the interaction between group and task.



**Discussion**

**Limitations:**

**Conclusions**

**References**

Annett, M. (1970). A classification of hand preference by association analysis. *British Journal of Psychology*, *61*(3), 303–321. https://doi.org/10.1111/j.2044-8295.1970.tb01248.x

de Cheveigné, A. (2020). ZapLine: A simple and effective method to remove power line artifacts. *NeuroImage*, *207*. https://doi.org/10.1016/j.neuroimage.2019.116356

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). Detection theory: A user’s guide. In *Routledge*. https://doi.org/10.4324/9781003203636

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., & Cornelissen, F. (2007). *What’s new in Psychtoolbox-3?* https://pure.mpg.de/rest/items/item\_1790332/component/file\_3136265/content

Koehn, P., & Monz, C. (2006). *Manual and Automatic Evaluation of Machine Translation between European Languages 1 Evaluation Framework*. 102–121. http://www.statmt.org/wmt06/

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/J.JCM.2016.02.012

MATLAB. (2018). *version 9.5.0.944444 (R2018b)*. The MathWorks Inc.

Pedroni, A., Bahreini, A., & Langer, N. (2019). Automagic: Standardized preprocessing of big EEG data. *NeuroImage*, *200*, 460–473. https://doi.org/10.1016/j.neuroimage.2019.06.046